Available online at http://www.ijasrd.org/in

International Journal of Advanced Scientific Research & Development

Vol. 02, Iss. 02, Ver. I, Apr – Jun' 2015, pp. 01 – 07

e-ISSN: 2395-6089
p-ISSN: 2394-8906

SUPERVISED ATTRIBUTE BASED CLASSIFICATION OF MICRO ARRAY SAMPLES

K.R. Saranya

Final Year M.E. Student, Department of CSE, Maha Barathi Engineering College, Chinnasalem, Tamil Nadu, India.

N. Sundaram

Assistant Professor, Department of CSE, Maha Barathi Engineering College, Chinnasalem, Tamil Nadu, India.

ARTICLE INFO

Article History:

Received: 16 May 2015;
Received in revised form:
19 May 2015;
Accepted: 19 May 2015;
Published online: 30 Jun 2015.

Key words:

Clustering,
Supervised,
Attributes,
Gene Expression.

ABSTRACT

For analyzing gene expression data there is a need in use of clustering Technique. Clustering provides alternative approach for heuristic for probability analysis. Classification and clustering are the two important ones in gene expression data analysis. Classification is concerned with assigning memberships to samples based on expression patterns, and clustering aims at finding new biological classes and refining existing ones. To cluster and recognize patterns in gene expression datasets, dimension problems are encountered. Typically, gene expression datasets consist of a large number of genes (attributes) but a small number of samples (tuples). In real data analysis, one of the important issues is computing both relevance and redundancy of attributes by discover the dependencies among them. Selection of genes is important in the clustering technique. So here it implements attribute clustering method which is able to group genes based on their interdependence so that meaningful patterns can be formed from the gene expression data. Grouping and selection are two important factors. We implement it by gene expression data so that meaningful clusters were formed.

INTRODUCTION

Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data. The clustering works on the principle of maximizing intra cluster and minimizing inter cluster similarities. The principle optimizes the overall function of the clustering process. Intra cluster similarities criterion ensure that objects or records that are placed in a group or cluster

How to cite this article: Saranyaa, K. R., & Sundaram, N., (2015). "Supervised Attribute Based Classification of Micro Array Samples". *International Journal of Advanced Scientific Research & Development (IJASRD), 02* (02/I), pp. 01 – 07.

have the highest similarities, while intra cluster similarities ensure that objects or records with in a cluster have a least similarity with those of another cluster. A clustering algorithm uses a similarity measure computed from the data to decide how closes the data points. Distance based measure is the most common similarity measure for quantitative data. The relationships between objects are represented by numerical values among different variable. These numerical representations, the data items, are stored in the dataset. Basically the analytical measurement process is applied to the dataset in order to group them. Measurement may be categorized in many ways; some of the distinctions arise from the nature of the properties based on the dataset.

1.1 Gene Expression Data

Gene Expression matrix is calculated from experimental procedure of scanning process; systematic variations arise from different combination of genes. The major problem of gene data set is missing value estimation and normalization of data according to gene similarity is important task. Many clustering approaches concern to determine the clusters in dataset and its required necessary preprocessing system. Filtering out genes at expression levels do not alter across samples, performing a logarithmic transformation at each expression level, or each row of gene expression matrix with mean of zero and variance of one need to preprocess the data set properly. Clustering dissimilar samples because of corresponding expression profiles may disclose sub cell types, which are hard to recognize by traditional morphology based approaches. Gene expression data is arranged in a data matrix, where each gene corresponds to one row and each condition to one column.

Gene expression matrices have been extensively analyzed in two dimensions: the gene dimension and the condition dimension. This analysis corresponds, respectively, to analyze the expression patterns of genes by comparing the rows in the matrix, and to analyze the expression patterns of samples by comparing the columns in the matrix.

1.2 Drawbacks in Existing Technologies

Microarray technology can simultaneously monitor the expression levels of thousands of genes during important biological processes and across collections of related samples. An important task of analyzing gene expression data is the detection of co expressed genes and coherent gene expression patterns. Coherent pattern is a “template,” while the expression profiles of the corresponding co expressed genes conform to the template with only small divergences. A microarray experiment often involves thousands of genes. However, only a small subset (perhaps several hundred) of those genes may play important roles in the underlying biological processes. The data set is decomposed into numerous small clusters. Some clusters will consist of groups of co expressed genes, while many clusters will be made up of intermediate genes.

Since there is no absolute standard, such as size or compactness, with which to rank the resulted clusters, it may require significant user effort to distinguish meaningful clusters from those trivial ones. Unsupervised similarity measures computed from the gene expressions, without using any information about the sample categories or response variables. The average expression level does not allow perfect discrimination of sample categories. Also, the existing algorithm avoids the noise sensitivity problem of existing unsupervised gene clustering algorithms.

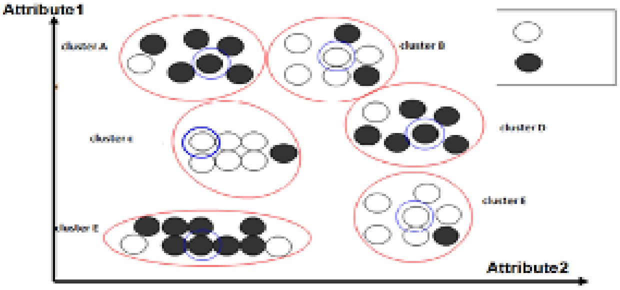
1.3 Supervised Attribute Clustering Algorithm

One of the important issues is computing both relevance and redundancy of attributes by discover the dependencies among them. The proposed supervised attribute clustering algorithm relies on mainly two factors, namely, Determining the relevance of each attribute and growing the cluster around each relevant attribute incrementally by adding one attribute after the other.

One of the important properties of the proposed clustering approach is that the cluster is augmented by the attributes those satisfy following two conditions:

1. Suit best into the current cluster in terms of a supervised similarity measure defined
2. Improve the differential expression of the current cluster most, according to the relevance of the cluster representative or prototype.

Fig 1: *Supervised Clustering*



RELATED WORKS

2.1 Cluster Analysis of Gene Expression Data

This uses the methods utilized in processing and analysis of gene expression data generated using DNA microarrays. Naturally, such an experiment requires computational and statistical analysis techniques. At the outset of the processing pipeline, the computational procedures are largely determined by the technology and experimental setup that are used. Subsequently, as more reliable intensity values for genes emerge, pattern discovery methods come into play. The most striking Peculiarity of this kind of data is that one usually obtains measurements for thousands of genes for only a much smaller number of conditions.

2.2 Grouping Selection and Classification of Gene Expression Data

The cluster and sub cluster of gene data set based on the resemblance matrix property of expression dataset. Clustering is the process of grouping data objects from a set of disjoint classes based on high similarity. It is an unsupervised classification, which does not rely on predefined classes and training examples. Thus, it is different from the pattern recognition or the area of statistics, which are referred to as discriminate analysis and assessment analysis.

2.3 Minimum Redundancy Feature Selection

Instead of using all available variables (features or attributes) in the data, one selectively chooses a subset of features to be used in the discriminate system. There are a number of advantages of feature selection: (1) dimension reduction to reduce the computational cost; (2) reduction of noise to improve the classification accuracy; (3) more interpretable features or characteristics that can help identify and monitor the target diseases or function types. However selecting a small subset out of the thousands of genes in microarray data is important for accurate classification of phenotypes. Widely used methods typically rank genes according to their differential expressions among phenotypes and pick the top-ranked genes. It observes that feature sets so obtained have certain redundancy and study methods to minimize it. This proposes a concept of minimum redundancy – maximum relevance (MRMR) feature selection framework

2.4 Mutual Information Based Selection

From an application-oriented point of view, an excessive input dimensionality implies lengthened preprocessing and recognition times, even if the learning and recognition performance is satisfactory. Looking into the application of the mutual criterion to evaluate a set of candidate features and to select an informative subset to be used as input data for a neural network classifier. Because the mutual information measures arbitrary dependencies between random variables, it is suitable for assessing the "information content" of features in complex classification tasks, where methods base on linear relations (like the correlation) are prone to mistakes. The fact that the mutual information is independent of the coordinates chosen permits a robust estimation. Nonetheless, the use of the mutual information for tasks characterized by high input dimensionality requires suitable approximations because of the prohibitive demands on computation and samples. An algorithm is proposed that is based on a "greedy" selection of the features and that takes both the mutual information with respect to the output class and with respect to the already-selected features into account.

2.5 Partial Least Squares

One important application of gene expression microarray data is classification of samples into categories. Modification of existing statistical methodologies or development of new methodologies is needed for the analysis of microarray data and propose a novel analysis procedure for classifying (predicting) human tumor samples based on microarray gene expressions. This procedure involves dimension reduction using Partial Least Squares (PLS) and classification using Logistic Discrimination (LD) and Quadratic Discriminate Analysis (QDA). Compare PLS to the well-known dimension reduction method of Principal Components Analysis (PCA). Under many circumstances PLS proves superior; illustrate a condition when PCA particularly fails to predict well relative to PLS.

SYSTEM DESIGN

It helps to identify functional groups of genes that are of special interest in sample classification and discrimination of sample categories. The supervised attribute clustering method uses this measure to reduce the redundancy among genes. It includes partitioning of the original gene set into some distinct subsets or clusters so that the genes within a cluster are tightly coupled with strong association to the sample categories. After forming the clusters, classifier is used to evaluate the accuracy of the generated clusters and feature subset selection. The supervised attribute clustering acts as an aid for microarray classification. Thereby it helps to increase the classification and predictive accuracy of the correlation based classifier. The proposed method reduces the dimensionality, avoids the noise sensitivity problem and increases the classification accuracy of microarray data and more number of infected cells can be found out that are unable to find out using classifier. Also it helps for early disease identification. The proposed system deals with different operations on the microarray data such as preprocessing, attribute clustering, classification and the performance evaluation with the existing system.

3.1 Preprocessing

Data pre-processing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values, impossible data combinations, missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis.

3.2 Evaluation

Identification of two types of gene selection such as occurrence based selection and sequence based selection. In occurrence based selection, provide the separate gene and show all gene which is provided by users. Then in sequence based selection, provide sequence and identify all sequences with position information.

3.3 Clustering

A new supervised attribute clustering algorithm is proposed to find co regulated clusters of genes whose collective expression is strongly associated with the sample categories or class labels. A new quantitative measure, based on mutual information, is introduced to compute the similarity between attributes.

3.4 Coherent Index Selection

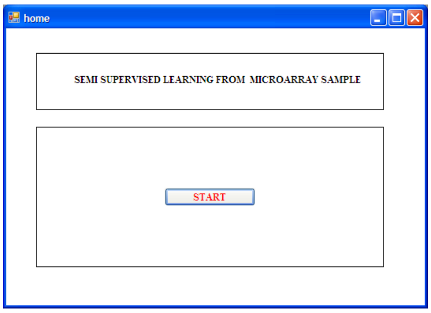
Calculate the gene positions. Then finding a good clustering method which contain interdependence information within clusters and discriminative information for classification. Selecting from each cluster significant genes with high multiple interdependence with other genes within each cluster. Also yielding very high classification results on both of gene expression datasets using a small pool of genes selected from the clusters found by as the training set.

3.5 Evaluation Criteria

Performance of the proposed supervised attribute clustering algorithm is extensively compared with that of some existing supervised and unsupervised gene clustering and gene selection algorithms. To analyze the performance of different algorithms, the experimentation is done on five microarray gene expression data sets. The major metrics for evaluating the performance of different algorithms are the class separately index and classification accuracy of naive baye's classifier, K-nearest neighbor rule, and support vector machine. To compute the classification accuracy, the leave-one-out cross validation is performed on each gene expression data set.

Here are the few implemented results;

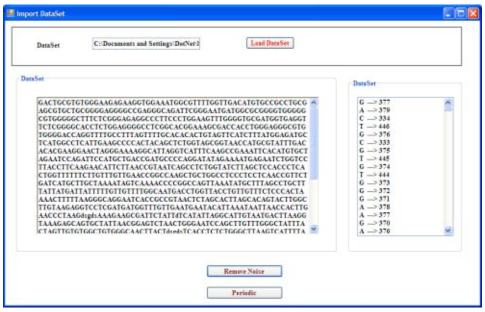
Main screen



home

SEMI SUPERVISED LEARNING FROM MICROARRAY SAMPLE

After loading microarray samples and segment dataset



Import DataSet

DataSet: C:\Documents and Settings\DotNet\

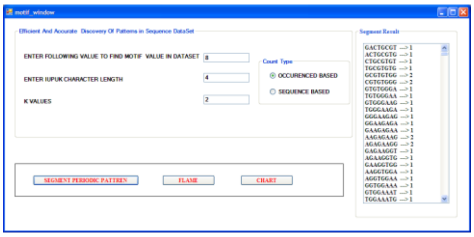
DataSet

```
GACTGCGTGTGGGAAGAGAAGGTGGAAATGGCGTTTTGGTTGACATGTGCCGCCTGCG
AGCGTGCTGCGGGGAGGGGCCGAGGGCAGATTCGGGAATGATGGCGCGGGGTGGGGG
CGTGGGGGCTTTCTCGGGAGAGGCCCTTCCCTGGAAGTTTGGGGTGCGATGGTGAGGT
TCTCGGGGCACCTCTGGAGGGGCCTCGGCACGGAAAGCGACCACCTGGGAGGGCGTG
TGGGGACCAGGTTTTGCCTTTAGTTTTGCACACACTGTAGTTCATCTTTATGGAGATGC
TCATGGCCTCATTGAAGCCCCACTACAGCTCTGGTAGCGGTAACCATGCGTATTTGAC
ACACGAAGGAACTAGGGAAAAGGCATTAGGTCATTTCAAGCCGAAATTCACATGTGCT
AGAATCCAGATTCCATGCTGACCGATGCCCCAGGATATAGAAAATGAGAATCTGGTCC
TTACCTTCAAGAACATTCTTAACCGTAATCAGCCTCTGGTATCTTAGCTCCACCCTCA
CTGGTTTTTTCTTGTTTGTTGAACCGGCCAAGCTGCTGGCCTCCCTCCTCAACCGTTCT
GATCATGCTTGCTAAAATAGTCAAAACCCCGGCCAGTTAAATATGCTTTAGCCTGCTT
TATTATGATTATTTTTGTTGTTTTGGCAATGACCTGGTTACCTGTTGTTTCTCCCACTA
AAACTTTTTAAGGGCAGGAATCACCGCCGTAACTCTAGCACTTAGCACAGTACTTGGC
TTGTAAGAGGTCCTCGATGATGGTTTGTTGAATGAATACATTAAATAATTAACCACTTG
AACCCTAAGdsgdsAAAGAAGCGATTCTATTdTCATATTAGGCATTGTAATGACTTAAGG
TAAAGAGCAGTGCTATTAACGGAGTCTAACTGGGAATCCAGCTTGTTTGGGCTATTTA
CTAGTTGTGTGGCTGTGGGCAACTTACTdsgdsTCACCTCTCTGGGCTTAAGTCATTTTA
```

DataSet

```
G ---> 377
A ---> 379
C ---> 334
T ---> 446
G ---> 376
C ---> 333
G ---> 375
T ---> 445
G ---> 374
T ---> 444
G ---> 373
G ---> 372
G ---> 371
A ---> 378
A ---> 377
G ---> 370
A ---> 376
```

Values for the clustering techniques



motif_window

Efficient And Accurate Discovery Of Patterns in Sequence DataSet

ENTER FOLLOWING VALUE TO FIND MOTIF VALUE IN DATASET: 8

ENTER IUPUK CHARACTER LENGTH: 4

K VALUES: 2

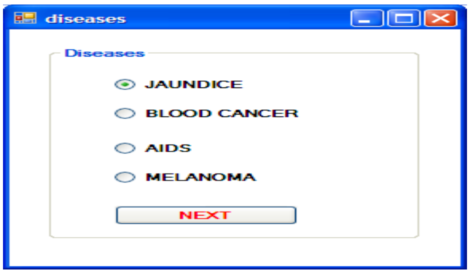
Count Type

- ● OCCURENCED BASED
- ○ SEQUENCE BASED

Segment Result

```
GACTGCGT ---> 1
ACTGCGTG ---> 1
CTGCGTGT ---> 1
TGCGTGTG ---> 1
GCGTGTGG ---> 2
CGTGTGGG ---> 2
GTGTGGGA ---> 1
TGTGGGAA ---> 1
GTGGGAAG ---> 1
TGGGAAGA ---> 1
GGGAAGAG ---> 1
GGAAGAGA ---> 1
GAAGAGAA ---> 1
AAGAGAAG ---> 2
AGAGAAGG ---> 2
GAGAAGGT ---> 1
AGAAGGTG ---> 1
GAAGGTGG ---> 1
AAGGTGGA ---> 1
AGGTGGAA ---> 1
GGTGGAAA ---> 1
GTGGAAAT ---> 1
TGGAAATG ---> 1
```

Select the input

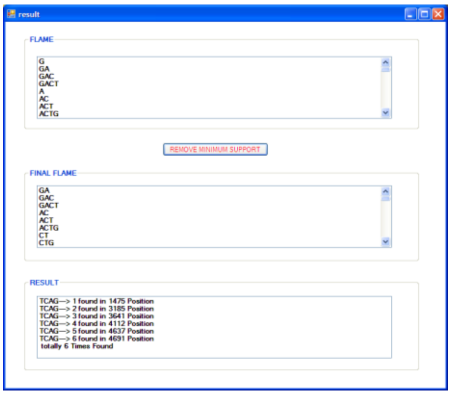


diseases

Diseases

- ● JAUNDICE
- ○ BLOOD CANCER
- ○ AIDS
- ○ MELANOMA

Calculation screen after the segmentation



result

FLAME

```
G
GA
GAC
GACT
A
AC
ACT
ACTG
```

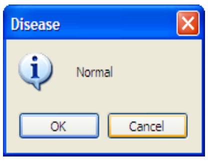
FINAL FLAME

```
GA
GAC
GACT
AC
ACT
ACTG
CT
CTG
```

RESULT

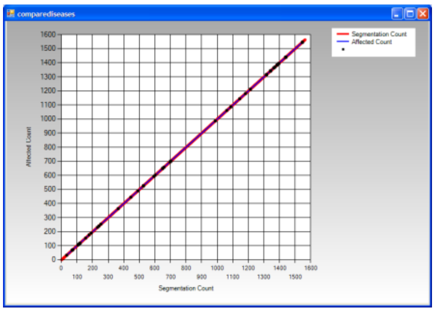
```
TCAG---> 1 found in 1475 Position
TCAG---> 2 found in 3185 Position
TCAG---> 3 found in 3641 Position
TCAG---> 4 found in 4112 Position
TCAG---> 5 found in 4637 Position
TCAG---> 6 found in 4691 Position
 totally 6 Times Found
```

Results



Normal

Comparison chart



CONCLUSION

The proposed supervised attribute clustering algorithm is based on measuring the similarity between attributes using the new quantitative measure, whereby redundancy among the attributes is removed. The clusters are then refined incrementally based on sample categories. The performance of the proposed algorithm is compared with that of existing supervised and unsupervised gene clustering and gene selection algorithms based on the class separately index and the predictive accuracy of naive baye's classifier, K-nearest neighbor rule, and support vector machine on three cancer and two arthritis microarray data sets.

REFERENCES

[1] M. Ankerst, M.M. Breunig, H.P. Kriegel, and J.Sander, “OPTICS: Ordering Points to Identify the Clustering Structure,” Proc. SIGMOD, pp. 49-60.

[2] Z. Bar-Joseph, E.D. Demaine, D.K. Gifford, N.Srebro, A.M. Hamel, and T.S. Jaakkola, “K-ary Clustering with Optimal Leaf Ordering for Gene Expression Data,” Bioinformatics, vol. 19, no. 9, pp. 1070- 1078, 2003.

[3] A. Ben-Dor, R. Shamir, and Z. Yakhini, “Clustering Gene Expression Patterns,” J. Computational Biology, vol. 6, nos. 3-4, pp. 281-297.

[4] M. Blatt, S. Wiseman, and E. Domany, “Super-Paramagnetic Clustering of Data,” Physical Rev. Letters, vol. 76, 1996

[5] Y. Cheng and G.M. Church, “Biclustering of Expression Data,” Proc. Eighth Int’l Conf. Intelligent Systems for Molecular Biology (ISMB), vol. 8, pp. 93-103, 2000.

[6] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, “Cluster Analysis and Display of Genome-Wide Expression Patterns,” Proc. Nat’l Academy of Sciences USA, vol. 95, no. 25, pp. 14863-14868.

[7] M. Ester, H. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” Proc. Second Int’l Conf. Knowledge Discovery and Data Mining, pp. 226-231.

[8] Pradipta Maji, “Mutual Information-Based Supervised Attribute Clustering for Microarray Sample Classification,” Science IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 1, January 2012.